Development and evaluation of a geographic information retrieval system using fine grained toponyms

Purves, Ross S ; Palacio, Damien ; Derungs, Curdin

Abstract: Geographic information retrieval (GIR) is concerned with returning information in response to an information need, typically expressed in terms of a thematic and spatial component linked by a spatial relationship. However, evaluation initiatives have often failed to show significant differences between simple text baselines and more complex spatially enabled GIR approaches. We explore the effectiveness of three systems (a text baseline, spatial query expansion, and a full GIR system utilizing both text and spatial indexes) at retrieving documents from a corpus describing mountaineering expeditions, centred around fine grained toponyms. To allow evaluation, we use user generated content (UGC) in the form of metadata associated with individual articles to build a test collection of queries and judgments. The test collection allowed us to demonstrate that a GIR-based method significantly outperformed a text baseline for all but very specific queries associated with very small query radii. We argue that such approaches to test collection development have much to offer in the evaluation of GIR.

DOI: https://doi.org/10.5311/JOSIS.2015.11.193

JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 11 (2015), pp. 1–29
doi:10.5311/JOSIS.2015.11.193

RESEARCH ARTICLE

Development and evaluation of a geographic information retrieval system using fine grained toponyms

Damien Palacio[1], Curdin Derungs[1,2], and Ross S. Purves[1]

[1]Department of Geography, University of Zurich, Switzerland
[2]URPP Language and Space, University of Zurich, Switzerland

Received: September 9, 2014; returned: January 9, 2015; revised: September 1, 2015; accepted: November 8, 2015.

Abstract: Geographic information retrieval (GIR) is concerned with returning information in response to an information need, typically expressed in terms of a thematic and spatial component linked by a spatial relationship. However, evaluation initiatives have often failed to show significant differences between simple text baselines and more complex spatially enabled GIR approaches. We explore the effectiveness of three systems (a text baseline, spatial query expansion, and a full GIR system utilizing both text and spatial indexes) at retrieving documents from a corpus describing mountaineering expeditions, centred around fine grained toponyms. To allow evaluation, we use user generated content (UGC) in the form of metadata associated with individual articles to build a test collection of queries and judgments. The test collection allowed us to demonstrate that a GIR-based method significantly outperformed a text baseline for all but very specific queries associated with very small query radii. We argue that such approaches to test collection development have much to offer in the evaluation of GIR.

Keywords: geographic information retrieval, toponyms, evaluation, user generated content

1 Introduction

Methods related to geographic information retrieval (GIR) [29] are key tools if we are to effectively explore and retrieve information with geographic context. The importance of geography in search generally was recently emphasized by White and Buscher [63]: "Since

studies have shown that one quarter of Web search queries have local intent (Himmelstein 2005), people's local experiences may bring significant benefit to others searching for local information [...]."

However, despite a number of papers which have demonstrated that a significant share of queries contain spatial information in some form (e.g., Excite [57], AOL [19], Yahoo [31]), and the seemingly intuitive conclusion that using geographic intelligence should improve query results in conjunction with the development of GIR, large scale evaluations (e.g., GeoCLEF [40]) have often failed to conclusively show the benefits of GIR over relatively simple text baselines. There are two potential explanations for this apparent disparity between an information need and evaluation. The first explanation is simply that geographic information is not special, and can in fact be processed using standard information retrieval (IR) methods (e.g., [3]), without recourse to approaches which handle it differently. However, an alternative reason for this mismatch may be that the properties of evaluation initiatives are not well suited to demonstrating the efficacy of GIR, for example, due to queries and document scopes of coarse spatial granularities or a relative lack of diversity in collections [8, 38]. The evaluation of GIR systems aiming to resolve queries containing fine grained, or local, information, applied to a corpus of detailed spatial information, is thus interesting for two reasons. Firstly, we may find support for White and Buscher's [63] hypothesis, namely that IR can significantly benefit from local information. Secondly, we can explore the strengths and weaknesses of different approaches to retrieval for a range of scenarios, and identify where GIR research efforts might be best concentrated.

Methods investigating retrieval at the level of fine spatial granularities are underrepresented in research into GIR, with many studies focusing on placenames at the granularity of towns and cities covering relatively large spatial extents (e.g., [8, 36]). Such coarse grained toponyms, especially within a given geographic region, appear to be less prone to ambiguity than those referring to smaller and less well known geographic features, such as mountains, hills, streams, or individual hamlets [6, 26]. Therefore, performing retrieval at finer local scales introduces an additional level of complexity. In previous work we proposed a disambiguation algorithm incorporating detailed geographic information available for all types of place names, independent of geographic feature type, size, or how well a place is known [14]. In a first, limited, evaluation we indicated that our disambiguation approach had the potential to outperform simple baseline approaches to toponym disambiguation [13] and in a pilot study suggested that user-generated content (UGC) could allow a more extensive evaluation of a complete GIR system [52].

Our aims in this paper are thus twofold:

- Firstly, we wish to explore the effectiveness [16] of a variety of implementations of a GIR system and a standard textual baseline in retrieving relevant documents from a corpus containing fine grained toponyms.
- Secondly, through these implementations, we wish to explore evaluation approaches more suited to GIR based on the use of UGC.

Crucially, this means that instead of using assessors to judge the relevance of documents to a query, we assume that metadata associated with a document by a user accurately summarizes the information content of a document. This approach allows us to semi-automatically build a very large test collection. Furthermore, since users associate coordinates and toponyms with content, it removes the problem of finding assessors with sufficient local knowledge to assess geographical content, which has previously been a

key stumbling block in the evaluation of GIR and related approaches at fine granularities [50, 63].

The remainder of this paper is organized as follows. In the next section, we review key background literature related to GIR, evaluation, and UGC and based on this review draw out a specific set of research challenges. We then describe the properties of our corpus, the GIR methods we used, and our approach to building a test collection for evaluation. Our results focus on the effectiveness of three approaches we took to document retrieval, and are discussed in the context of the aims set out above.

2 Background

Our review is focused on three distinct, but linked, research areas. Firstly, we give a brief review of key steps in GIR, and some of the challenges therein. Secondly, we give an overview of evaluation in IR, and more particularly GIR, setting out steps required to measure retrieval effectiveness. Thirdly, we look at the phenomenon of UGC, and more specifically its use in evaluation, with a focus on content related to geography.

2.1 GIR

A commonly accepted definition of GIR is that proposed by Jones and Purves [29] who state that "GIR is therefore concerned with improving the quality of geographically-specific information retrieval with a focus on access to unstructured documents such as those found on the web."

Three elements of this definition are particularly important and worthy of emphasis here. Firstly, the definition emphasizes "unstructured documents"—textual documents without any explicit semantic mark-up. Secondly, "geographically-specific" retrieval is key—there is an expectation that both queries and results are evaluated with respect to this geography. Thirdly, by relating GIR to information retrieval, the authors implicitly adopt both methods and measures from the IR literature. Although others in GIScience have placed more emphasis on retrieval using richer underlying and derived semantics than is often the case in (G)IR (e.g., work from Nedas and Egenhofer [48] on spatial-scene similarity queries or Ballatore et al. [4] on semantic similarity) such approaches, to the best of our knowledge, have not been embedded within complete GIR systems. Where such a system is implemented, a number of key challenges arise [29] including the detection and disambiguation of geographical references from text, spatial indexing, retrieval, and ranking of documents for a given query, and approaches to evaluating effectiveness.

Identifying and associating toponyms which occur in text with unique geographic locations (and thus metric coordinates) is a key initial step in the processing of corpora. Any approach must be capable of dealing with a special case of the more general problem of word sense disambiguation [47]—identifying the meaning of individual words in context. With respect to toponyms two problems are particularly important: so-called *geo/geo*, ambiguity (e.g., is the London being referred to in Canada or the UK) and *geo/non-geo* ambiguity (e.g., does Turkey refer to a country or a bird). Garbin and Mani [20] reported that 40% of all toponyms occurring in an average text had more than one possible referent location, while Leveling and Veiel [37] found geo/non-geo ambiguity for up to 17% of terms identified as

candidate toponyms in newspaper articles. Correctly assigning unique spatial references to a text therefore requires that both forms of ambiguity be automatically dealt with.

Navigili [47] points out that a key problem in word sense disambiguation is the generation of additional knowledge which can be used to provide additional context in the disambiguation process. This observation holds true for toponym disambiguation, where typical approaches use a variety of contextual information, to derive disambiguation rules (e.g., SPIRIT [60], PIV [18], DIGMAP [43]) or train machine learning algorithms [44]. Contextual information usually incorporates toponym-related information often available at the granularity of cities or towns, such as population, administrative hierarchies, feature types or physical areas. At finer spatial granularities, for instance, toponyms referring to small or less well known geographic features such as hills, rivers or individual hamlets, this information is usually not available or relevant (e.g., population) and toponym ambiguity appears to be more pronounced [6,26]. Consequently, many state of the art disambiguation approaches are not well suited to geoparsing detailed spatial descriptions containing local information since additional contextual information is not available.

Having identified individual toponyms in a text, and associated them with unique coordinates, it is possible to index documents for retrieval. Indexing techniques using terms contained in documents are well established in IR [3]. Typically, documents are converted into an inverted file structure, i.e., a list of terms, each associated with documents containing this term. In GIR spatial indexes, such as quadtrees, can be built using a collection of spatial document footprints. These footprints can take different forms, including a collection of points (representing the coordinates associated with all toponyms resolved in the document), a single point (representing the mean center of all toponyms), or more complex structures such as bounding boxes, convex hulls or even density maps (e.g., [17, 60]).

Having built an index, retrieval is concerned with returning a ranked list of documents for a given query (e.g., "hiking in Zermatt"), where document rank is related to a similarity score between the query and each document. Commonly used and very effective baseline text methods incorporate relative frequencies of query terms within documents compared to overall frequencies in the corpus as a whole (e.g., TF-IDF, BM25, etc. [3]). Geographic similarity scores are often approximated by spatial metrics—for example, computing the relative overlap of a spatial query and the spatial footprints of documents [17, 34]. GIR approaches can then rank results for a given query using independent textual and spatial indexes and combine them (e.g., [60]), where the method used for combination from text and spatial indexes is of particular importance.

It is possible to combine documents according to scores [35] or rankings [46] and approaches using both families of technique have previously been shown to be effective [51]. Figure 1 illustrates how CombMNZ (which uses scores) and Borda (which uses rankings) combine results lists for a given query q, for the example query "Matterhorn hiking," where we assume that the query results are generated by a system using textual and spatial indexes as described above. Here, thematic results (a) are returned for the criterion "hiking" and spatial results (b) for the criterion "Matterhorn." For each document (d_i) we have a rank (r), and a computed similarity score between the query q and the document d_i (s).

2.2 Evaluation

In traditional IR many large scale evaluation campaigns, such as TREC [62] or CLEF [5] have followed the Cranfield model (see Sanderson [56] for a detailed review of test collec-

r	d	s
1	d_4	14.5
2	d_3	12
3	d_5	8.7
4	d_2	0.5
5	d_1	0.5

(a) Thematic Ranks and Scores

r	d	s
1	d_6	150
2	d_1	120
3	d_4	80
4	d_2	30

(b) Spatial Ranks and Scores

↓ ↓

Combination of result lists

↓ ↓

r	d	Combined similarity s					
1	d_4	2.83	=	$2 \cdot (\frac{14.5-0.5}{14.5-0.5}$	+	$\frac{80-30}{150-30})$
2	d_1	1.50	=	$2 \cdot (\frac{0.5-0.5}{14.5-0.5}$	+	$\frac{120-30}{150-30})$
3	d_6	1.00	=	$1 \cdot (0$	+	$\frac{150-30}{150-30})$
4	d_3	0.82	=	$1 \cdot (\frac{12-0.5}{14.5-0.5}$	+	$0)$
5	d_5	0.58	=	$1 \cdot (\frac{8.7-0.5}{14.5-0.5}$	+	$0)$
6	d_2	0.00	=	$2 \cdot (\frac{0.5-0.5}{14.5-0.5}$	+	$\frac{30-30}{150-30})$

(c) Combination with normalized CombMNZ

r	d	Combined similarity s
1	d_4	$8 = 5 + 3$
2	d_6	$5 = 0 + 5$
3	d_1	$5 = 1 + 4$
4	d_2	$4 = 2 + 2$
5	d_3	$4 = 4 + 0$
6	d_5	$3 = 3 + 0$

(d) Combination with Borda count

r: rank, d: document, s: similarity between the query and a document d_i

Figure 1: Illustration of result lists combination with (c) normalized CombMNZ versus (d) Borda count [51].

tion based evaluation in IR). These campaigns allowed comparison of the effectiveness of different retrieval approaches and systems, and this model was also adopted in GIR where, for example, the GeoCLEF initiative specifically explored the effectiveness of geographic search [40].

The following three components make up the basics of a test collection [25, 56]:

- **A corpus**: The corpus is a collection of documents. For example, the GeoCLEF 2007 corpus consisted of three sub-corpora in English, German, and Portuguese, each with around 200,000 documents [41].
- **A set of topics (also known as queries)**: Topics are generated in accordance with the available information from the corpus. Typically, only topics which one might expect to be answerable using the corpus are incorporated in the collection. Topics in GIR often consist of a theme, a spatial relationship, and a location (e.g., "Hiking (theme) near (spatial relationship) Zermatt (location) "). Usually, a minimum of 25 topics are considered to be necessary to allow testing of statistical significance between different configurations [61].
- **Query relevance judgments (qrels)**: Qrels are a measure of whether documents in the corpus are judged by assessors as relevant to a given query. In most cases it is not possible to assess qrels for all possible query-document combinations and a pooling approach is used, such that top ranked documents for each query are merged to create a pool of documents which are then judged.

The evaluation of a GIR system for effectiveness is very time consuming. The (maximum[1]) number of assessor judgments that must be collected is:

$$\textit{assessor judgments} = \textit{approaches} \times \textit{components} \times \textit{queries} \times \textit{top documents} \times \textit{judgements per document} \quad (1)$$

For instance, Purves et al. [53] evaluated the SPIRIT system by comparing two approaches with three components (two components, plus the combination) on the basis of 38 queries associated with the top ten results and collected a total of 456 judgments from two annotators. GeoCLEF 2008 [40], the largest known evaluation initiative in GIR, contains judgments for a total of approximately 600,000 English, Portuguese, and German documents. Despite the effort of manually annotating such large numbers of documents, the GeoCLEF collection had a relatively narrow topical focus and almost exclusively consisted of newspaper articles. Furthermore, queries were developed and proposed without having detailed knowledge of the spatial properties of the collection. Thus, typical queries had a rather coarse geographic granularity and the collection is unsuitable for evaluation of a GIR approach optimized for fine grained spatial information.

Although GeoCLEF provided a potential starting point, no widely used GIR test collections have to date been developed. One important reason is probably that although individual investigations incorporating methods from GIR often demonstrated improvements over IR baselines (e.g., [9, 51, 53]), large-scale evaluations, such as GeoCLEF, often failed to show the same improvements [40]. Possible reasons for this mismatch include the coarse spatial granularity of the topics and text corpus, consisting of newspaper articles and serious difficulties in judging geographic relevance, especially at finer granularities [8, 11, 38, 50]. The lack of a widely acknowledged GIR test collection is one reason why evaluations of GIR systems are often omitted [39], only focus on retrieval efficiency, such as indexing time or storage use [60], or are limited in extent [51, 53].

The effectiveness of IR and GIR systems respectively, is often measured by reporting accepted measures including precision and recall [42]. Table 1 summarizes some of the main evaluation measures. While precision is commonly reported, recall is often omitted since it requires knowledge of all possible relevant documents for each query (which in turn implies judging the relevance of every document in a corpus for every query). The two measures generally have opposite trends such that a high precision comes at the cost of low recall and vice versa. For this reason, precision and recall are often combined in a single measure, such as average precision (AP). AP incorporates the ranking of each retrieved document and thus results in a global score for each query. APs for all queries can be averaged to give a mean average precision (MAP). For example, MAP ranged from 0.21 to 0.28 in GeoCLEF 2007 [41] and from 0.27 to 0.3 in GeoCLEF 2008 [40]. Although Moffat and Zobel [45] report on shortcomings of average precision (and by extension MAP) due to its dependence on estimated recall values and recall's apparent lack of correspondence to user satisfaction, it continues to be a very commonly used measure in both GIR and IR. Furthermore, we suggest that where precision is low, recall and thus average precision values become much more relevant in understanding search effectiveness, especially where complete knowledge of a corpus is possible.

[1] "Maximum" since often results from different approaches or components overlap and thus need not be judged twice.

Measure	Formula
Precision	$P(s,q) = \frac{Rel(q) \cap Ret(s,q)}{Ret(s,q)}$
Recall	$R(s,q) = \frac{Rel(q) \cap Ret(s,q)}{Rel(q)}$
Average Precision	$AP(s,q) = \frac{\sum_{r=1}^{Ret(s,q)} P(s,q,r) * is\_rel(r)}{Rel(q)}$
Mean Average Precision	$MAP(s) = \frac{\sum_{q=1}^{Nb_q} AP(s,q)}{Nb_q}$

$Rel(q)$: number of relevant documents for the query q,
$Ret(s,q)$: number of documents retrieved by the system s for the query q,
Nb_q : number of queries
$is\_rel(r)$: binary function which returns 1 if the result is relevant

Table 1: Example evaluation measures.

2.3 Evaluation, crowd sourcing, and user generated content

Crowd sourcing has recently been adopted as a rapid and cost-effective approach to gathering large numbers of relevance judgments in IR. For example, Alonso and Mizzaro [1] used Amazon's Mechanical Turk to replicate TREC experiments, while Sanderson et al. [58] used crowd sourced judgments to compare evaluation measures to user satisfaction. However, a key difference to the scenarios presented in these papers concerns the geographic component of relevance in GIR. Previous experience has shown that annotators have more difficulty, and disagree more often when assessing geographic rather than thematic relevance (e.g., [11, 50]). Thus, although crowd sourcing appears to be a very effective way of judging thematic topic relevance, we suggest that its potential for judging geographic relevance, especially at fine granularities is essentially unclear.

Parallel to the rise of crowd sourcing as an approach to evaluation, researchers have also become increasingly interested in the use of user generated content (UGC) as a knowledge source. UGC often consists of a primary piece of information, such as photographs in Flickr or a short text messages in Twitter and secondary information added either by a user or the system. For example, in Flickr, secondary information consists of tags added by users and timestamps and GPS coordinates automatically generated by the system. Researchers using UGC often assert that the information reflects non-expert user concepts allowing, for example, folksonomies to be built (e.g., [28]). Approaches with a geographic focus, often make the underlying, implicit, assumption that UGC reflects some form of naïve geographical knowledge [15]—information with semantic content not available from other, official data sources reflecting more vernacular or lay uses of terms and place names. For instance, Rattenbury et al. [54], used Flickr to investigate the spatio-temporal distribution of tags, which they argued were a rich source of place semantics, while Grothe and Schaab [23] used Flickr to explore the use of vague place names. Within GIScience, much research has focused on the quality of UGC, which has often been shown to be surprisingly high. Haklay et al. [24] demonstrated that OpenStreetMap was spatially very accurate (with digitized information on average within 6m of data from a national mapping agency), although coverage was incomplete, while Hollenstein and Purves [27] found that Flickr images were both accurately and precisely annotated in terms of coordinates and toponyms associated with individual instances (for example, 86% of images tagged with Hyde Park were found in or around Hyde Park). Such observations are taken advantage of in machine learning

approaches, where geo-referenced data is used to train and evaluate classifiers seeking to assign coordinates to documents, for instance in the form of Wikipedia articles or Twitter feeds [33, 64, 66].

Within IR, metadata contributed by users has been used in a variety of contexts. Thus, in medical IR so called ICD codes provided by clinicians are commonly used in evaluation tasks (e.g., [32]). However, to the best of our knowledge, though UGC has been used to train and evaluate individual components of systems (e.g., georeferencing as described above) it has not been used to evaluate a complete GIR system. This is to some extent surprising since, as we suggest above, evaluating geographic relevance appears to be a particularly challenging task, especially at finer granularities. Indeed, we would argue that although metadata attached by UGC authors represents everyday concepts, this is not non-expert knowledge, as is often claimed in other UGC related contexts. On the contrary, we argue authors are probably the individuals best suited to associating their descriptions with metadata on the topic or important waypoints, and indeed in the case of descriptions of hikes have likely experienced the place they are describing, in contrast to assessors judging the relevance of documents for a given query. However, it is important to emphasize that before using such metadata in evaluation exercises, further assessment of data quality (cf. [24, 27]) is required.

In summary, we identify the following research challenges which we wish to address in the remainder of this paper:

- Toponym disambiguation has mostly focused on coarse spatial granularities. Local information and detailed spatial descriptions remain a challenge for GIR.
- GIR has often not outperformed simple IR baselines using simple text indexes for resolving spatial queries, especially in large scale evaluations.
- Compilation of (GIR) test collections remains very time consuming, and because of lack of local knowledge, difficult.
- The potential quality of UGC as an alternative means of evaluating GIR, and the necessary properties of UGC collections used in this context, remain unclear.

3 Materials and methods

In the following, as laid out in Figure 2, we describe four key elements of the work carried out in this paper. These are:

(1) our **corpus** of UGC which contains both **metadata** and textual **descriptions**;
(2) q **test collection**, consisting of a set of documents, queries, and relevance judgments for each query and document;
(3) **GIR approaches**, capable of returning ranked lists of documents from the test collection, for any given query; and
(4) an **evaluation**, testing the retrieval effectiveness of the different GIR approaches with respect to our test collection.

3.1 UGC corpus

We chose to use a UGC corpus describing outdoor activities in this paper, for a number of reasons. Firstly, we expect that those undertaking such activities are likely to be familiar

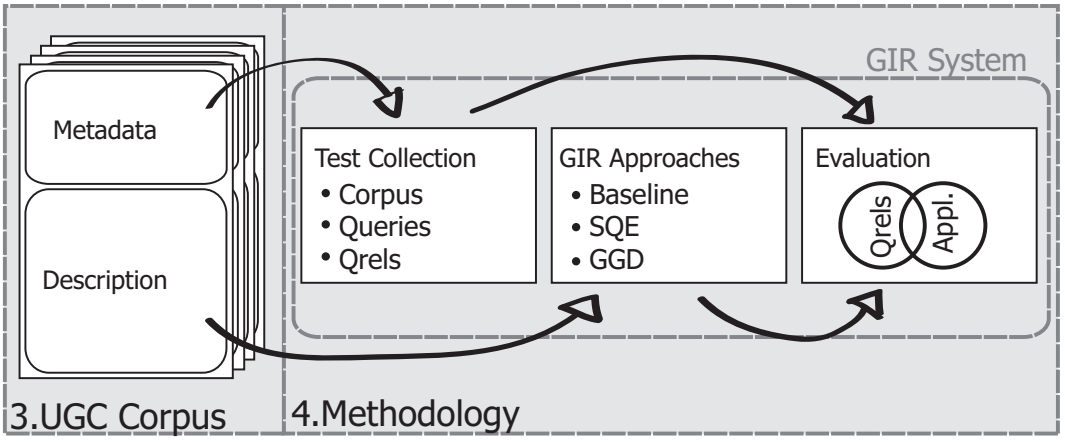

Figure 2: Overall framework of the experiments described in this paper.

with the use of geographic information, and thus to accurately assign metadata to their reports. Secondly, examination of text describing hikes and other outdoor activities shows that the use of fine grained toponyms is very common. We mined documents from the Hikr homepage (http://http://www.hikr.org) where each document consists of a header, containing metadata, and textual descriptions, usually in the form of a detailed report of a trip lasting one or more days. The trips can all be considered outdoor activities with different thematic foci, ranging from biking or hiking to alpine mountaineering.

We collected a total of 58,803 documents. Of these, 85% (50,000) have at least one type of activity (theme) and one waypoint (coordinates) listed in the metadata. About 57% of the documents describe activities in Switzerland and 68% of all descriptions are in German (12% in Italian and 3% in French). The metadata and the description are usually contributed by the same author, however, we cannot control for this. The descriptions consist of unstructured text and have a median length of 258 words (ranging from 1 to 4000). 66% of all descriptions are between 100 and 500 words (i.e., one to two pages), whereas only 3% are more than 1000 words in length. The authors of the descriptions use usernames and are thus anonymous. The distribution of Hikr documents over authors is uneven, as often reported in the UGC literature (e.g., [49]), such that 1% of all registered users (n=10,000) write approximately 90% of all descriptions.

3.1.1 Metadata

The metadata consists of a regional classification imposed by Hikr, the date, information relating to the activity, its difficulty, and waypoints. We used the two metadata fields "activity" and "waypoints." The activity field distinguishes nine themes, namely hiking, biking, mountaineering, skiing, climbing, snowshoe hiking, ski touring, ice climbing, and via ferrata, with an over-representation of descriptions of hiking trips (Figure 3). Waypoints are added as a list of hand annotated toponyms for each document, associated with coordinates (i.e., no toponym ambiguity). Hikr maintains a gazetteer of toponyms used in the metadata. Most of these toponyms overlap with well known toponyms from official

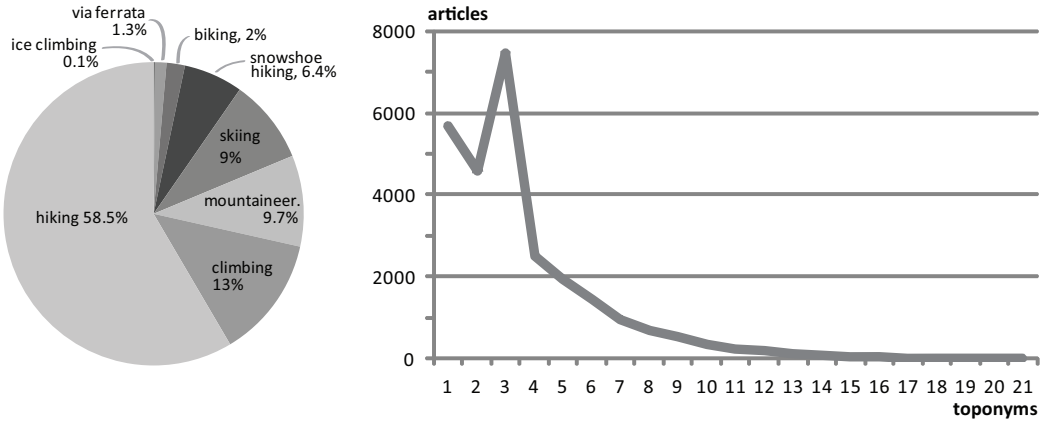

Figure 3: Metadata info on activities and waypoints per document.

gazetteers, but some refer to vernacular place names, such as the names of climbing routes or local landmarks.

This metadata, and in particular the waypoints and activities, is of central importance to us, since we consider them as groundtruth with respect to the spatial and thematic context of the respective textual descriptions. On average metadata for one article contains 3.62 toponyms (ranging from 1 to 72). 66% of documents have between 1 and 3 toponyms and only 6% have more than 8 (Figure 3, right). In total the Hikr metadata contains a total of 97880, and 15630 unique, toponyms, and each document is annotated with an average of 1.28 activities, probably because mountaineering or climbing routes often commence with a hike. Nonetheless, only 4% of all documents contain more than 2 activities.

3.1.2 Data quality

Since we intended to use the corpus both to generate a test collection and test different approaches, it was important to document the properties of both the metadata and the descriptions. To better understand the quality of the metadata, we first undertook a simple analysis of a random set of 50 documents, and explored the use of waypoints within these. We hypothesized that typical tours would have a goal (e.g., a peak), a start, and potentially a different end point, as well as intermediate locations considered important by users. We then classified waypoints according to a simple schema. Of our 50 documents 49 had an explicit goal, which in 44 cases was a mountain peak and for the remaining five a mountain hut. 27 documents also mentioned a start point, and of these 24 also listed an end point. No tours listed an end point in the absence of a start and 25 documents also listed other intermediate waypoints. Finally, the median number of waypoint types was three. These results are important, as they clearly demonstrate that users typically add metadata which describes different aspects of the route and capture some notion of spatial spread.

Having ascertained that metadata existed which appeared to characterize routes, the next question which can be asked concerns the quality of this metadata. Since waypoints in Hikr are associated with user-defined locations, and since we also had access to an au-

thoritative gazetteer of Swiss toponyms (SwissNames), we explored the overlap between these two sources for the complete collection. In total, around 69% of the 16,634 toponyms present as Hikr waypoints were also found in the SwissNames gazetteer as exact textual matches. Since users also entered coordinates for these toponyms, we then examined the distribution of Euclidean distances between the authoritative SwissNames gazetteer and user-entered Hikr waypoints (Figure 4). These results demonstrate that the quality of geolocation is high, with more than 96% of toponyms positioned by users within a distance of less than 1km from authoritative SwissNames data. This has important implications, since it means not only are the toponyms correctly positioned, but that it is reasonable to treat the 31% of toponyms which are not exact matches with Swissnames data as reliable, additional gazetteer information.

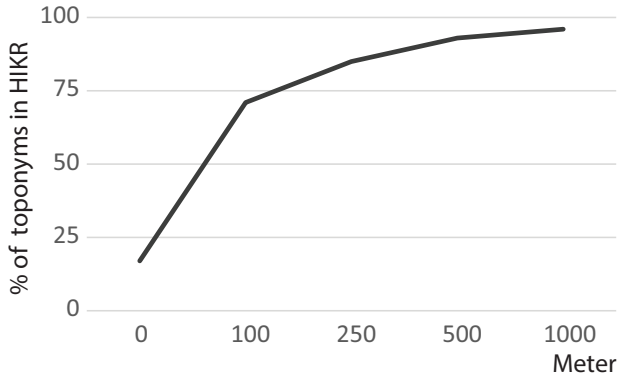

Figure 4: Cumulative curve of distance between Hikr toponyms and their entries in the authoritative SwissNames gazetteer.

Having ascertained the quality of the metadata, the next question posed is the relationship of this metadata to our descriptions, and to the toponyms that our system identified. Here we are interested in two key aspects. Firstly, how similar are the toponyms used in the metadata to those we identified in the text and secondly, how representative are the coordinates of toponyms found in the metadata of the information we extracted from the textual description. In general, we expect textual descriptions to contain more toponyms than the metadata (based on the above analysis for a random subset of metadata). This relationship is confirmed, for the complete dataset, in Figure 5a, where the boxplot shows that the median number of waypoints as a function of toponyms identified in the text is 0.47. The scatterplot (Figure 5b) underpins this analysis with most, but not all, points lying above the line indicating a 1:1 relationship between metadata waypoints and found toponyms. The final question that can be posed is how representative are the waypoints given of the toponyms identified in the descriptions. Since, typically, fewer waypoints exist than toponyms we calculated the nearest neighbor distance to a waypoint for every toponym identified in our descriptions. This measure captures the accuracy of the data better than a simple centroid, since it snaps found toponyms to the most appropriate waypoints, and reflects the spread of our data. The histogram in Figure 5c clearly shows that these distances are in general small, and in fact the median distance from waypoints to identified toponyms was 819m for 201,000 toponyms.

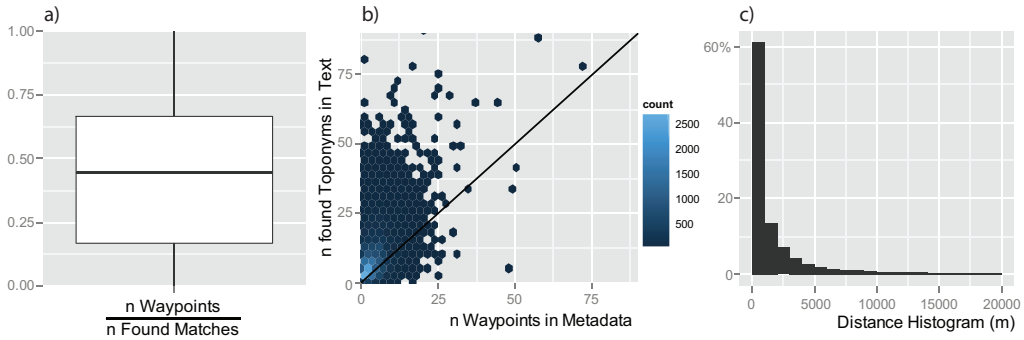

Figure 5: a: Box plot of ratio between waypoints in metadata and toponyms found in text documents; b: relationship between waypoints in metadata and toponyms found in text documents; c: histogram of distances between nearest waypoint and toponyms found in each document.

Our analysis of data quality suggests that users have, at a minimum, attributed routes with toponyms which match textual descriptions and assigned the correct coordinates to these. Furthermore, the metadata appears to give a reasonable summary of the locations of the toponyms found in the descriptions. Thus, it appears reasonable to assume that the spatial metadata can be used as ground truth information, and we henceforth assume that the waypoints in the metadata contain explicit information on relevant spatial locations locating the textual description.

3.2 GIR approaches

In the following we describe the three approaches we took to performing GIR on the descriptions of the Hikr documents. These incorporated a simple textual baseline, spatial query expansion, and an approach optimized for resolving fine spatial granularity information. Geographic queries took the form of a thematic topic (e.g., "hiking") and a spatial element, expressed as a set of coordinates. This can be seen as analogous to a typical search interface using a map interface, where a user specifies a location through a map click, and the topic of interest by entering terms in a text box. Furthermore, regions were specified by associating a radius with a location.

The **text-based baseline system** performed retrieval [22] using BM25 for document ranking [55]. To generate a query the toponym nearest to the location expressed by coordinates was found in a gazetteer, and combined with a topic to generate a query (e.g., "hiking Zermatt"). Although this may seem simplistic, such simple text baselines have often been shown to outperform more sophisticated GIR systems especially for containment queries (e.g., [21, 41, 53]).

A second approach applied **spatial query expansion (SQE)** for the spatial part of the query. All toponyms found within the query radius were used to retrieve a set of documents from the search index, ranked using BM25. As query radius increases, so does the number of toponyms associated with the query. In general, it is assumed that spatial query expansion may be effective in finding additional relevant documents by adding additional

toponyms to the initial query [59] without the need for the overhead of a spatial index, and, since combinations of toponyms are more likely to be unique than individual toponyms, to reduce problems with toponym ambiguity. A second ranked set of documents was retrieved from the index using the thematic part of the search as a query, before results were combined using one of three methods: strict intersection (i.e., relevant documents must appear in both sets), CombMNZ, or Borda.

The third approach [13], **geometric geomorphometric disambiguation (GGD)** is optimized for retrieving geographic information from textual descriptions, independent of spatial granularity, such that it is of particular use for text descriptions containing local information. In contrast to the other two approaches described, toponyms in the descriptions are identified, disambiguated, and assigned unique coordinates which are then associated with document IDs and stored in a spatial index. The basic approach firstly performs toponym lookup and, secondly, applies toponym disambiguation (Figure 6). Toponym lookup is a simple comparison between all words (and up to five word combinations) in the Hikr descriptions and entries in a list of toponyms complied by combining toponyms found in the Hikr metadata and the Swissnames gazetteer. The result of the toponym lookup is a set of candidate toponyms for each document of which most are ambiguous. Unambiguous toponyms are used as anchor points in the disambiguation process which incorporates two measures. For the combination of each ambiguous toponym and all anchor point toponyms from the same description, we compute the mean Euclidean distance and the mean topographic similarity. Both means are weighted using the distance in words between the corresponding toponyms in text. Topographic similarity is computed using a measure that we introduced and tested in previous work [14] which compares morphological properties of the candidate locations using a digital elevation model. The underlying assumption is that an ambiguous toponym can be disambiguated with appropriate confidence if it is either proximate to other unambiguous toponyms or has similar topographic properties (e.g., it is also a mountain).

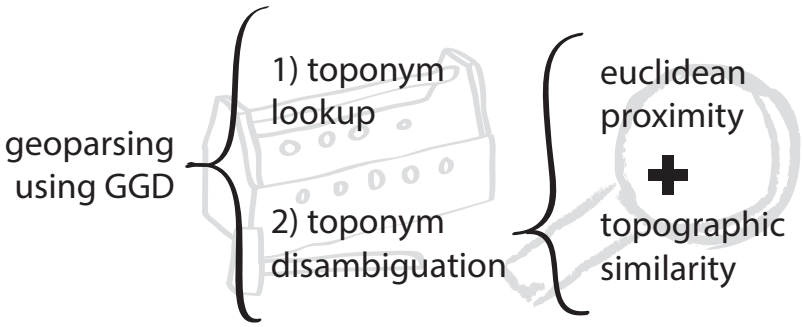

Figure 6: Sketch of the components and the functionality of GGD.

All disambiguated toponyms are then stored in a spatial index represented as an R-tree in a PostGIS database. Using the spatial index, we can process the spatial parts of queries using spatial intersection. For a given query location and search radius we retrieve all documents with one or more toponym locations within the radius centred on the query location and bounded by the query radius. The topical query again takes the form of a

simple term search, ranked using BM25, as for spatial query expansion, and results were combined from the two indices again using strict intersection, CombMNZ, or Borda.

3.3 Test collection

The most important difference between our approach and typical state of the art evaluations of GIR systems is that we use a UGC corpus, namely Hikr to generate both topics and query relevance judgments (qrels). Importantly, the qrels are based on metadata provided for every document, meaning that we have a complete set of metadata. In the following description, we are particularly interested in how such metadata can be used to improve the evaluation of GIR, and thus focus on these aspects. The steps are not restricted to Hikr and could be applied to any UGC where text descriptions are available in parallel with metadata.

The evaluation was carried out using the 26,974 Hikr documents written in German (language recognition was performed using a language-detection library[2]) found in Switzerland. This metadata was used as groundtruth and, given the results of our quality assessment, we assumed that waypoints contain explicit information relating to relevant spatial locations described in textual descriptions. With respect to the thematic component of queries, we restricted our analysis to the activity classification given in the Hikr metadata. Our queries thus took the form of a topic (e.g., "hiking") and a point coordinate. This point coordinate was associated with the nearest toponym found in SwissNames, to give a query of the form "hiking in Zermatt." Each query thus incorporates "where" and "what" characteristics, or spatial and topical dimensions respectively, and we focus on the simplest spatial relationship (containment), where GIR systems have often struggled to outperform simple IR baselines (e.g., Purves [53]).

Articles were judged thematically relevant if they were annotated with the appropriate topic (e.g., hiking) in their metadata and spatially relevant if they were associated with waypoints found proximal to the query location. Proximity was approximated by a set of search radii, namely 1, 2, 5, 10km. The minimum radius of 1km is greater than the median distance (819m) we found between found toponyms and waypoints, which thus represents a minimal granularity for our system. Increasing radii can be associated with different spatial information interests, for example, a radius of 1km will deliver local information for a very precise region, whereas larger spatial radii are interesting for those who want to discover a region, perhaps on foot (e.g., 5km), or by bicycle (10km). In contrast to the nine activity topics that are used in metadata and thus considered in the queries, we incorporate a very large number of spatial queries, distributed over the Swiss mountains.

To compile a list of queries containing both spatial and thematic information, which were used in the test collection and for the evaluation, we combined all nine themes with approximately 1600 spatial queries (grid points) and the four search radii. For the baseline textual search, grid points were assigned the nearest toponym from a gazetteer. This results in a set of some 56.000 queries. Qrels were computed for every query-document combination separately, such that we know which documents are relevant for which queries. A document is relevant only if both thematic and spatial information in the metadata match with the specifications in the query. This also means that for each of the four search radii to be tested, the result set for the spatial queries increases (increasing search radii typically increases the number of documents spatially relevant). We only retained queries in our

[2]http://http://code.google.com/p/language-detection/

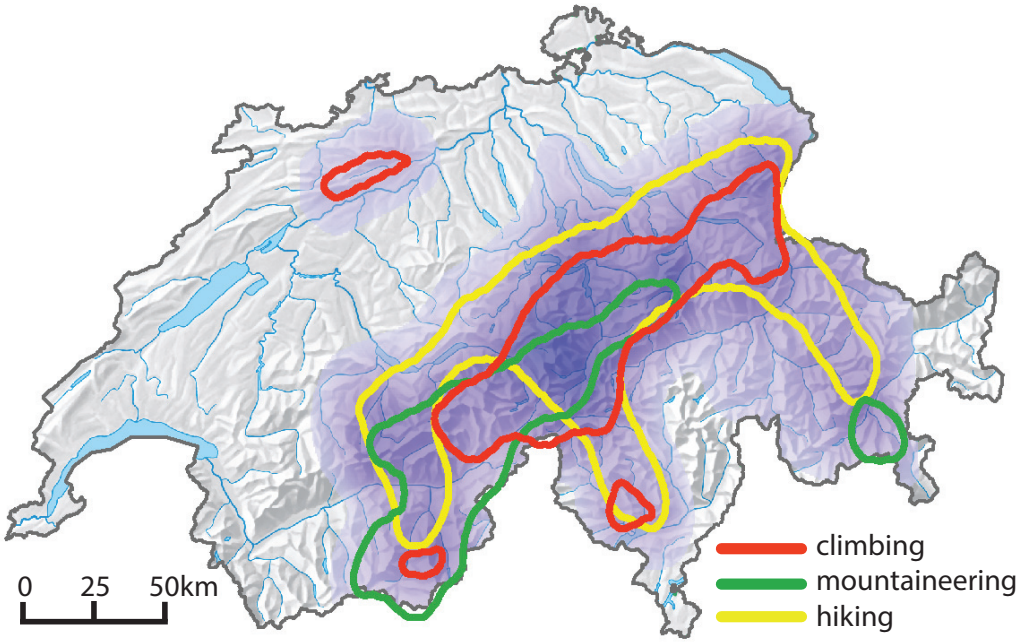

Figure 7: Spatial representation of all queries (n=4354). The blue background color shows the density of all spatial queries, with the three contour lines identify the 80% densest regions for different topics.

collection if at least ten relevant documents existed (i.e., documents whose thematic and spatial information matched the query). Filtering all query combinations for this criterion we retained a final list of 4354 queries. This is a very large number of queries, compared to the minimum number of queries typically required for statistical testing (i.e., 25). Due a small number of relevant documents per query, two themes (snowshoeing and ice climbing) were discarded.

Figure 7 is a density map of all spatial queries, where only queries where at least ten relevant documents were available for retrieval were retained in order to generate representative results. The density map of all the queried locations thus shows the spatial footprint of the Hikr corpus and simultaneously maps three outdoor activities in Switzerland, with a clear focus on German speaking regions in the Swiss Alps. The densest regions of each of the three topics largely overlap with the density computed from all spatial queries. The spatial particularities of each topic region match our expectation, such as for instance the density peak for climbing in the northwest of Switzerland, overlapping with the limestone cliffs of the Jura mountains and the generally more extensive region associated with hiking, extending into the foothills of the Alps.

4 Results and interpretation

In this section we describe the comparison of the three approaches as discussed in Section 3.2 applied to the test collection as introduced in Section 3.3. The test collection consisted of a set of some 26,000 individual documents and around 4400 queries and relevance judgments, gathered from the metadata, for each document-query combination.

We focus here on the evaluation measures precision, recall, average precision (AP) and mean average precision (MAP) as well as a set of precision-recall curves (cf. Table 1). Table 2 shows MAP values for the three approaches and all four search radii, applied to the purely spatial queries, while Figure 8 shows box plots of precision (P@10, P@20 and P@100) and average precision (AP), for each of the four search radii.

We can make a number of observations based on these results:

- Spatially-intelligent methods (SQE und GGD) outperform IR-baselines for all cases and do so statistically significantly for all search radii except 1km. For very local results (ca. 1km) performance of all methods is similar, probably because at this search radius we approach the underlying granularity of the toponym data and the quality of our ground truth has a similar granularity (median difference between toponyms in text and waypoints was 819m).
- GGD outperforms SQE statistically significantly for all search radii except 1km.
- The difference between the simple textual baseline, interpreting locations as terms, and the other two approaches is most obvious for larger search radii (5–10km), where the baseline has precisions much lower than those for either SQE or GGD. The text baseline is thus not suited to retrieving results relevant for larger regions.

Approach	1km	2km	5km	10km
Baseline	0.31	0.21	0.09	0.05
SQE	0.35	0.39^{**}	0.47^{**}	0.50^{**}
GGD	$\mathbf{0.35}^{*}$	$\mathbf{0.42}^{**}_{\dagger\dagger}$	$\mathbf{0.55}^{**}_{\dagger\dagger}$	$\mathbf{0.63}^{**}_{\dagger\dagger}$

Table 2: MAP (reported to two decimal places) for spatial queries. Symbols * and ** denote a significant difference (respectively $p < 0.05$ and $p < 0.01$, Students T-test) compared with Baseline. Similarly symbol † denotes a significant difference with SQE.

4.1 Multicriteria queries

The retrieval results for multicriteria queries (i.e., consisting of a spatial and thematic component), such as "Hiking in Zermatt," vary according to how the results are combined. The results for the three approaches previously introduced, using GGD (the best performing method in our spatial search) and compared across the four search radii are summarized in Table 3.

MAP for Borda is statistically and practically significantly higher compared to the other two approaches when spatial and thematic result sets are combined. Therefore, we use only Borda for intersection in the next section, where the results of the three approaches (i.e., baseline, SQE and GGD) are compared for spatial and thematic queries.

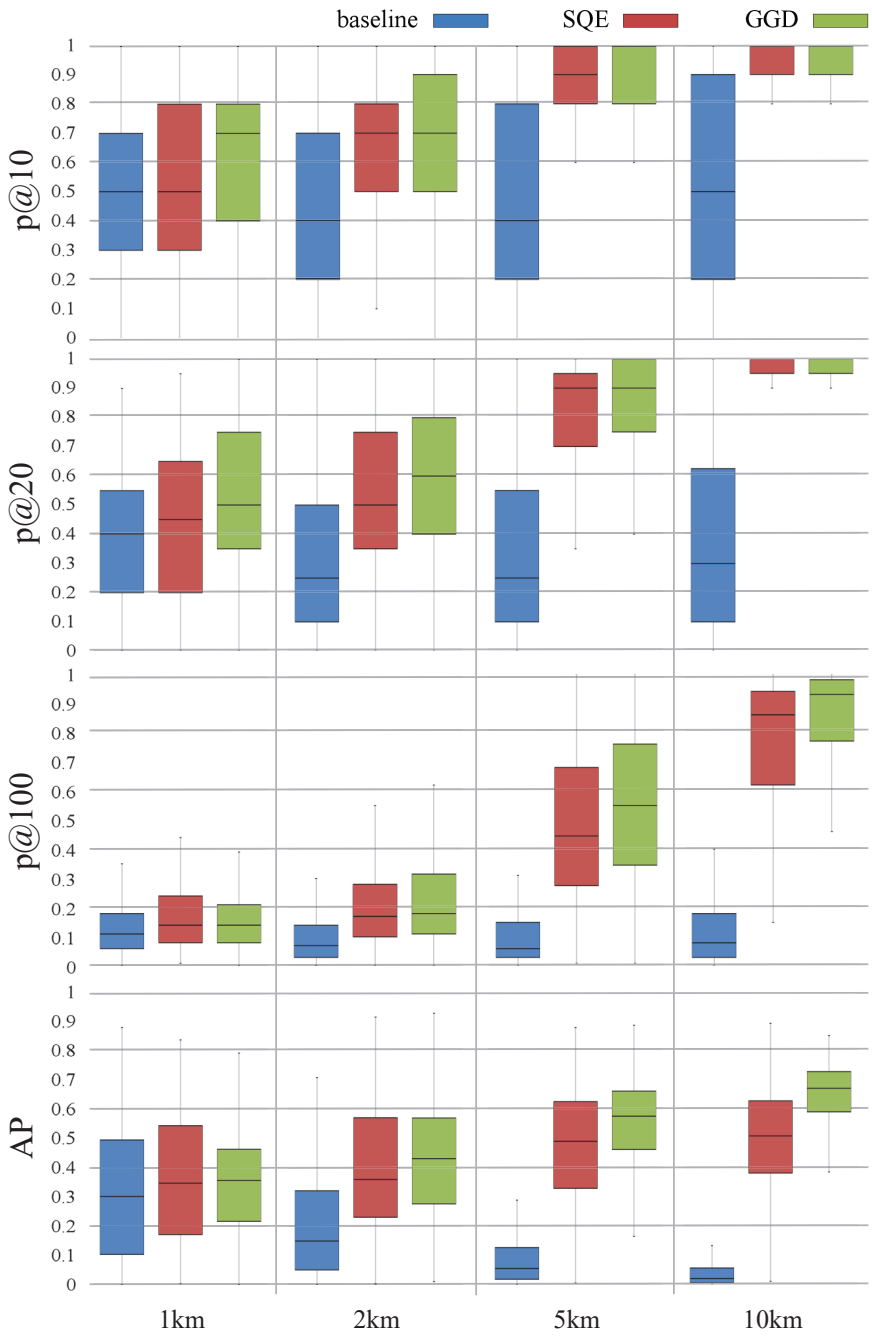

Figure 8: Summary statistics for P@10, P@20, P@100 and AP for spatial queries using four different search radii and three GIR approaches.

Approach	1km	2km	5km	10km
Intersection	0.11	0.12	0.13	0.14
CombMNZ	0.11	0.12	0.13^{**}	0.15^{**}
Borda	$\mathbf{0.22}^{**}_{\dagger\dagger}$	$\mathbf{0.23}^{**}_{\dagger\dagger}$	$\mathbf{0.22}^{**}_{\dagger\dagger}$	$\mathbf{0.24}^{**}_{\dagger\dagger}$

Table 3: MAP (reported to two decimal places) for three results lists combinations applied to GGD. Symbol ** denotes a significant difference ($p < 0.001$, Students T-test) compared with Intersection. Similarly symbol †† denotes a significant difference with CombMNZ.

4.2 Spatial and thematic queries

The evaluation in this section is analogous to that of a typical GIR system, comparing both a thematic and spatial component for a simple spatial relationship (in), and we investigate the performance of three configurations of our system, namely a baseline, and two contrasting GIR approaches: SQE and GGD, on the basis of 4354 queries containing spatial and thematic information. Table 4 illustrates MAP values for the three approaches and all four search radii, applied to spatial and thematic queries. GGD outperforms the other two approaches for all but the 1km search radius. The differences between the results for larger search radii are practically and statistically significant. At a search radius of 1km the baseline text method outperforms the two GIR methods, though the difference with GGD is not statistically significant.

Approach	1km	2km	5km	10km
Baseline	**0.25**	0.16	0.06	0.03
SQE	0.14^{**}	0.10^{**}	0.09^{**}	0.11^{**}
GGD	$0.22_{\dagger\dagger}$	$\mathbf{0.23}^{**}_{\dagger\dagger}$	$\mathbf{0.22}^{**}_{\dagger\dagger}$	$\mathbf{0.24}^{**}_{\dagger\dagger}$

Table 4: MAP (reported to two decimal places) of GIR approaches. Symbol ** denotes a significant difference ($p < 0.001$, Students T-test) compared with Baseline. Similarly symbol †† denotes a significant difference with SQE.

Figure 9 illustrates the results as box plots for precision (P@10) and average precision (AP), for each of the four search radii.

A clear advantage of the complete knowledge associated with our corpus is the range of measures which can be calculated. This is perhaps best illustrated by the precision-recall curves shown in Figure 10. Here the differences in performance between the three implementations are clearly illustrated. The textual baseline's performance decays rapidly with increasing search radius and recall, while precision for SQE, though more stable with increasing search radii, also decays rapidly as recall increases. By contrast, GGD produces results where precision decreases more slowly with very little variation in performance as a function of search radius.

Based on the results in Table 4 and Figures 9 we deduce the following:

- A simple text baseline gives the best performance at a radius of 1km, similar to that of GGD. One possible reason is the relative rarity of toponyms in comparison to the themes, which in turn gives these a higher weight in the ranking in the textual baseline and allows this method to retrieve a reasonable number of relevant documents

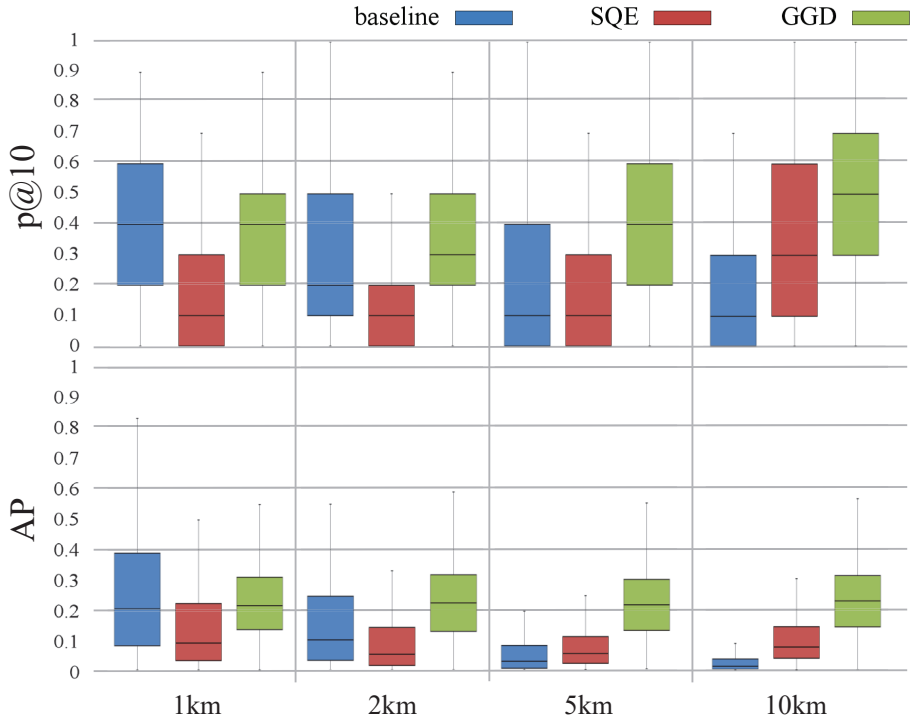

Figure 9: Summary statistics of P@10 and AP for spatial and topical query results combined using Borda for four search radii.

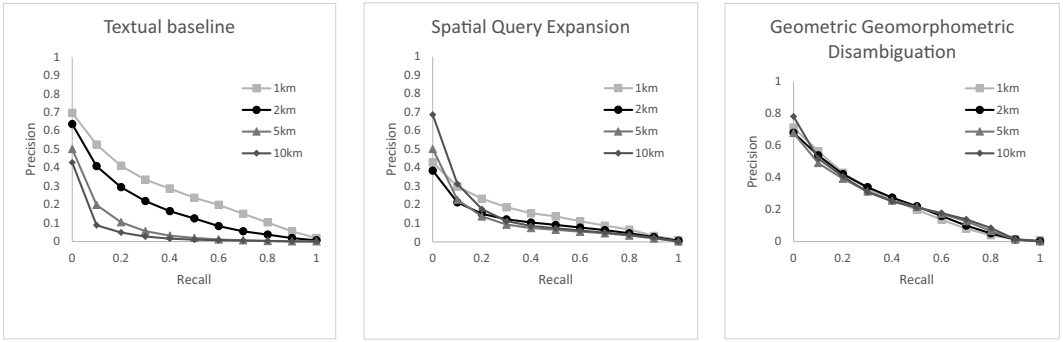

Figure 10: Precision-recall curves for the three GIR approaches.

(though it is important to note that P@10 is still only of the order of 0.4, implying that 6 out of 10 top-ranked retrieved documents were not relevant).
- In general P@10 and AP values for fine granularity queries containing spatial and thematic information are quite low, ranging between 0.08 and 0.52 (P@10) and 0.05 to 0.30 (AP).
- GDD always outperformed SQE, and was as good as, or better than the textual baseline for all search radii.
- Performance of the text baseline decreases with increasing search radii, as was the case for the purely spatial search. This again demonstrates that the text baseline performs well only when the (spatial) search granularity is similar to the underlying granularity of the toponym data.
- SQE in general performs poorly in a spatial and thematic search. This suggests that any gain in recall from the extra toponyms proposed is offset by a resulting decrease in precision due to ambiguity.
- The method used to combine results from different indexes is important: we found Borda (based on document rank) to be most effective.
- GGD shows very stable performance across a range of search radii.

The results represented in Table 4 and Figure 9 are well suited to indicating general trends. However, by exploring the results for individual queries it is possible to gain insight into some of reasons for differences in performance. Table 5 lists six queries and the precisions obtained for each of the three methods applied.

#	Thematic query	Spatial query	Baseline	SQE	GGD
1	Skiing	Wachtlammstock	0.0	0.4	0.7
2	Climbing	Cavigliano	0.1	0.5	0.5
3	Via ferrata	Libige	0.1	0.5	0.2
4	Hiking	Fellilücke	0.6	0.0	0.1
5	Via ferrata	Bargis	0.7	0.6	0.6
6	Mountaineering	Loch	0.0	0.0	0.0

Table 5: Precision values for six individual queries and the three approaches.

In the following, based on an analysis of the individual documents retrieved and the properties of the corpus as a whole, we discuss the properties of each method for all six queries:

#1 The theme "skiing" is prominent in the corpus, however, the toponym "Wachtlammstock" is not explicitly mentioned in descriptions and thus the baseline does not retrieve any relevant documents. GGD results in high precision values since it incorporates the intersection between the spatial index of documents and the search region. SQE does not perform as well since many of the toponyms neighboring "Wachtlammstock" are ambiguous and thus irrelevant documents are retrieved.

#2 Again, although the theme occurs in the corpus, the toponym "Cavigliano" is rare and thus the text baseline has poor performance. In contrast to the previous query, the toponyms neighboring Cavigliano are less ambiguous, and SQE performs similarly to GGD.

#3 The results for this query are similar to #1 and #2, with the important difference that GGD performs poorly as a result of errors in the toponym disambiguation process.

#4 Hiking is the most prominent theme in the corpus. At the same time, the combination with "Fellilücke" is rare and documents with this combination of terms are highly likely to be relevant, and thus the baseline performs well. Both SQE (because of nearby ambiguous toponyms) and GGD (because of errors in disambiguation) perform poorly.
#5 All approaches work well, mainly because "Via ferrata" and "Bargis" only rarely occur in the corpus and if they do, they occur in combination. In this case it is not necessary to use spatial intelligence or a sophisticated disambiguation. Simple text queries (baseline) or query expansion (SQE) will do the job.
#6 All approaches fail in retrieving relevant documents for this query. The reason is that "Mountaineering" is a prominent topic, whereas "Loch" is highly ambiguous, with several referent locations in Switzerland (geo/geo ambiguity) and occurrence as a common word in general language (geo/non-geo ambiguity). In principle, good disambiguation (through GGD) should improve results, but here this is clearly not the case.

In summary, it is clear that exploring the properties of individual queries allows us to better understand the properties of different implementations, and suggest where efforts might best be invested in tuning the system.

5 Discussion and conclusions

As we set out in the introduction, this paper had two key aims: firstly to test effectiveness of a range of configurations of a GIR system and a textual baseline in retrieving relevant documents using fine grained toponyms and, secondly, to assess the efficacy of a UGC collection in performing such an evaluation and make some general recommendations for GIR evaluation on this basis. In so doing, we also wished to address a number of research gaps including the following: evaluation of individual GIR systems has often relied on relatively small collections of user judgments (e.g., [53]), in larger scale evaluation efforts GIR has often not been found to outperform traditional IR (e.g., [40]), and GIR has often focused on resolving toponyms with relatively coarse spatial granularities (e.g., [36]).

We explored three basic system configurations: a purely text-based IR approach, which concatenated a toponym and a theme to build a query, a GIR-baseline, using query expansion to generate multiple toponyms within a given search radius, and a system (GGD) designed to disambiguate toponyms and query using a combination of a spatial and textual index. Both spatially enhanced methods retrieved two sets of results and combined these using either strict intersection or methods based on relevance scores or ranks.

We initially evaluated individual components of the system, namely purely spatial search, and the influence of different methods to combine rankings before exploring the performance of all three configurations for standard GIR queries such as "Hiking in Zermatt." Here we focus on the results for the full GIR queries, noting that the evaluation suggested that a text baseline was likely to significantly decrease in performance for large search radii, and that Borda was the most appropriate way of combining results for our system. The GGD-based GIR system outperformed purely textual search for all query radii greater than 1km. This result has two key implications, which we believe also have general implications for GIR.

Firstly, purely textual search for small query radii performed as well or better than spatially enhanced methods (e.g., MAP=0.25 as opposed to 0.22 for GGD and 0.14 for SQE). This result is in line with previous work (cf. [40]) indicating the challenge of outperforming a simple baseline, especially for containment queries (e.g., [53]). We would argue that there are two underlying reasons for this, at first glance perhaps surprising, result.

(1) None of the three systems is perfect, and toponym disambiguation at these fine spatial granularities is very challenging. For small query radii, choosing the right toponym is more or less as effective as choosing the right coordinates since both are equally specific.
(2) As is well known to linguists and those studying spatial cognition, toponyms are very efficient ways of communicating locations in natural language [7,12]. Since our system has full knowledge about toponyms at a very fine granularity, it can choose appropriate queries in the form of toponyms for a given coordinate - something that may be much harder for a human user of a system with incomplete knowledge.

An important implication of these results is that identifying representative toponyms [65] may be an effective way of supporting spatial search without recourse to full GIR. Clearly, toponyms also have different granularities, and knowledge of the region related to a toponym stored in a gazetteer may be a very useful attribute in such tasks [30].

Furthermore, as search regions increase, GGD, not only outperforms both the text baseline and SQE, but shows very stable performance. Where GGD has correctly disambiguated toponyms in a document (and it is important to note that GDD uses multiple toponyms iteratively to add documents to the spatial index), the method stands a good chance of finding thematically relevant documents. Search radii has little or no influence on the performance of GGD, since documents are treated as "bags of points." By contrast, baseline text performance becomes very poor for larger radii, since a single toponym typically fails to adequately capture information about the region. SQE's performance is more stable than the text baseline, but also significantly worse than that of GGD, demonstrating that, at least for fine grained toponyms, this method is not effective in disambiguation. Thus, our results clearly show a case where the overhead of building a spatial index is worthwhile, even for relatively straightforward containment queries. This is the first time we are aware of a significant increase in performance by a GIR-based system for containment queries. We believe this has two reasons. Firstly, typical approaches to evaluation have not allowed experiments of this nature, in particular varying query radii to be carried out, and thus are probably most comparable to our 1 km search radius (where the performance of GGD and the text baseline are broadly equivalent). Secondly, we are also not aware of previous experiments with such fine grained toponyms, where accurate disambiguation becomes increasingly important as the influence of ambiguity increases [6].

In this paper we used what may be seen as a very specialized corpus, consisting of both unstructured text (descriptions of hikes) and structured metadata (containing a variety of information including coordinates and classification of activities) to build a test collection. Our evaluation was performed on some 26,000 documents, for which queries containing coordinates (or associated toponyms for purely textual search), a radius, and an activity were created. By selecting only queries for which at least ten relevant documents existed, we then created a total of 4354 queries for evaluation, a much larger set than we are previously aware of in GIR evaluation (cf. [40]). For every query, a complete set of document relevance judgments was available, allowing us to calculate a wide range of evaluation

measures including precision and recall. This approach means that, particularly for the spatial part of queries, we remove the problems of judging relevance based on the use of typically ambiguous, and very detailed toponyms. In previous work we found that judging toponyms automatically assigned to images was very challenging, even for individuals with local knowledge, at fine spatial granularities [50], and that interannotator agreements were generally better for what have been termed generic queries (e.g., "is this document about mountains") as opposed to specific (e.g., "does this document refer to Schwarzhorn near Davos, one of many mountains of that name") [2].

Nonetheless, our approach has a number of disadvantages. The most significant of these concerns the thematic part of our queries. Queries were generated using one of six themes assigned in the metadata. Thus, this part of the system has a relatively straightforward task (effectively text retrieval using one of six possible terms). However, increasing the number of thematic queries, and gathering appropriate relevance judgments, could in principle be carried out using crowd sourcing (e.g., [1]), since such generic parts of queries appear to be easier to consistently judge. Furthermore, we rely on the locations suggested by users as being relevant to a textual description, which are typically a very small number of point locations often referring to some form of linear geometry (in the form of a trip in the mountains). However, our detailed, and we would argue indispensable, quality analysis demonstrated that not only were the metadata provided of high quality, but also in good agreement with the textual descriptions given.

Based on our experiences, we would argue that the time has come for researchers focusing on GIR to re-evaluate the need for shared resources in evaluation. Rather than generating relevance judgments TREC style, we propose that the use of UGC metadata associated with pure text is a sensible way forward. Although such corpora are less common than, for example, news corpora as previously used in GIR, other areas where such data is available (e.g., using georeferenced Tweets to train classifiers [66] or benchmark data to develop methods to locate images and videos [10]) suggests that joint resources can rapidly accelerate progress. In the context of GIR we propose that identifying and pooling resources such as that identified in this paper (of which we believe there are many more than are currently acknowledged) may be a productive way of advancing current research efforts. We suggest that there may be real value in returning to previous evaluation efforts and trying to better characterize the geographic properties of corpora (for example, by identifying and mapping all toponyms found), such that query effectiveness can be better understood. Our results provide clear evidence that GIR methods can improve the effectiveness of search for local information - but also demonstrate the challenges in developing test collections appropriate for this task.

Acknowledgments

The research reported in this paper was funded by the project FolkOnt supported by the Swiss National Science Foundation under contract 200021-126659. We also thank three anonymous reviewers for their constructive suggestions which helped us improve and clarify this paper. Elise Acheson is also thanked for her helpful comments on the manuscript.

References

[1] ALONSO, O., AND MIZZARO, S. Using crowdsourcing for TREC relevance assessment. *Information Processing & Management 48*, 6 (Nov. 2012), 1053–1066. doi:10.1016/j.ipm.2012.01.004.

[2] ARMITAGE, L. H., AND ENSER, P. G. Analysis of user need in image archives. *Journal of Information Science 23*, 4 (1997), 287–299. doi:10.1177/016555159702300403.

[3] BAEZA-YATES, R., AND RIBEIRO-NETO, B. *Modern Information Retrieval*. ACM Press, New York, 1999. doi:10.1.1.27.7690.

[4] BALLATORE, A., WILSON, D. C., AND BERTOLOTTO, M. Computing the semantic similarity of geographic terms using volunteered lexical definitions. *International Journal of Geographical Information Science 27*, 10 (2013), 2099–2118. doi:10.1080/13658816.2013.790548.

[5] BRASCHLER, M., AND PETERS, C. Cross-language evaluation forum: Objectives, results, achievements. *Information Retrieval 7*, 1-2 (Jan. 2004), 7–31. doi:10.1023/B:INRT.0000009438.69013.fa.

[6] BRUNNER, T. J., AND PURVES, R. S. Spatial autocorrelation and toponym ambiguity. In *Proc. Geographic Information Retrieval (GIR)* (2008), pp. 25–26. doi:10.1145/1460007.1460013.

[7] BURENHULT, N., AND LEVINSON, S. C. Language and landscape: A cross-linguistic perspective. *Language Sciences 30*, 2 (2008), 135–150. doi:10.1016/j.langsci.2006.12.028.

[8] CARDOSO, N. Evaluating geographic information retrieval. *SIGSPATIAL Special 3*, 2 (July 2011), 46–53. doi:10.1145/2047296.2047307.

[9] CHEN, Y.-Y., SUEL, T., AND MARKOWETZ, A. Efficient query processing in geographic web search engines. In *Proc. ACM SIGMOD International Conference on Management of Data* (Chicago, IL, 2006), ACM Press, pp. 277–288. doi:10.1145/1142473.1142505.

[10] CHOI, J., THOMEE, B., FRIEDLAND, G., CAO, L., NI, K., BORTH, D., ELIZALDE, B., GOTTLIEB, L., CARRANO, C., PEARCE, R., AND POLAND, D. The placing task: A large-scale geo-estimation challenge for social-media videos and images. In *Proc. 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia* (New York, NY, 2014), ACM, pp. 27–31. doi:10.1145/2661118.2661125.

[11] CLOUGH, P., JOHO, H., AND PURVES, R. Judging the spatial relevance of documents for GIR. In *Proc. 28th European Conference on IR Research (ECIR)* (Berlin, 2006), vol. 3936 of *Lecture Notes in Computer Science*, Springer, pp. 548–552. doi:10.1007/11735106_62.

[12] DAVIES, C., HOLT, I., GREEN, J., HARDING, J., AND DIAMOND, L. User needs and implications for modelling vague named places. *Spatial Cognition & Computation 9*, 3 (2009), 174–194. doi:10.1080/13875860903121830.

[13] DERUNGS, C., PALACIO, D., AND PURVES, R. S. Resolving fine granularity toponyms: Evaluation of a disambiguation approach. In *Proc. 7th International Conference on Geographic Information Science (GIScience)* (Colombus, USA, 2012), pp. 1–5. doi:10.5167/uzh-67546.

[14] DERUNGS, C., AND PURVES, R. S. From text to landscape: Locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science 28*, 6 (2014), 1272–1293. doi:10.1080/13658816.2013.772184.

[15] EGENHOFER, M. J., AND MARK, D. M. Naive geography. In *Proc. Conference on Spatial Information Theory (COSIT)* (Berlin, 1995), Lecture Notes in Computer Science, Springer, pp. 1–15. doi:10.1007/3-540-60392-1_1.

[16] FRØKJÆR, E., HERTZUM, M., AND HORNBÆK, K. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proc. SIGCHI Conference on Human Factors in Computing Systems* (The Hague, 2000), vol. 2, ACM, pp. 345–352. doi:10.1145/332040.332455.

[17] FRONTIERA, P., LARSON, R., AND RADKE, J. A comparison of geometric approaches to assessing spatial similarity for GIR. *International Journal of Geographic Information Science 22*, 3 (2008), 337–360. doi:10.1080/13658810701626293.

[18] GAIO, M., SALLABERRY, C., ETCHEVERRY, P., MARQUESUZAA, C., AND LESBEGUERIES, J. A global process to access documents' contents from a geographical point of view. *Journal of Visual Languages And Computing 19*, 1 (2008), 3–23. doi:10.1016/j.jvlc.2007.08.010.

[19] GAN, Q., ATTENBERG, J., MARKOWETZ, A., AND SUEL, T. Analysis of geographic queries in a search engine log. In *Proc. First International Workshop on Location and the Web (LocWeb)* (New York, NY, 2008), ACM, pp. 49–56. doi:10.1145/1367798.1367806.

[20] GARBIN, E., AND MANI, I. Disambiguating toponyms in news. In *Proc. Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (Vancouver, Canada, 2005), Association for Computational Linguistics, pp. 363–370. doi:10.3115/1220575.1220621.

[21] GEY, F. C., LARSON, R. R., SANDERSON, M., JOHO, H., CLOUGH, P., AND PETRAS, V. GeoCLEF'05: The CLEF 2005 cross-language geographic information retrieval track overview. In *Proc. 5th Workshop on Cross-Language Evalution Forum (CLEF)* (Berlin, 2006), vol. 4022 of *Lecture Notes in Computer Science*, Springer, pp. 908–919. doi:10.1007/11878773/101.

[22] GOSPODNETIĆ, O., AND HATCHER, E. *Lucene in Action*. Manning Publications, 2005.

[23] GROTHE, C., AND SCHAAB, J. Automated footprint generation from geotags with kernel density estimation and support vector machines. *Spatial Cognition & Computation 9*, 3 (2009), 195–211. doi:10.1080/13875860903118307.

[24] HAKLAY, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design 37*, 4 (2010), 682–703. doi:10.1068/b35097.

[25] HARMAN, D. K. The TREC test collections. In *TREC: Experiment and Evaluation in Information Retrieval*, E. M. Voorhees and D. K. Harman, Eds. MIT Press, 2005, ch. 2, pp. 21–53.

[26] HILL, L. L. *Georeferencing: The Geographic Associations of Information*. MIT Press, Cambridge, UK, 2006. doi:10.1080/00330120701787274.

[27] HOLLENSTEIN, L., AND PURVES, R. Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science 1*, 1 (2010), 21–48. doi:10.5311/JOSIS.2010.1.3.

[28] HOTHO, A., JÄSCHKE, R., SCHMITZ, C., AND STUMME, G. Information retrieval in folksonomies: Search and ranking. In *Proc. 3rd European Conference on The Semantic Web: Research and Applications (ESWC)* (Berlin, 2006), Springer, pp. 411–426. doi:10.1007/11762256_31.

[29] JONES, C. B., AND PURVES, R. S. Geographical information retrieval. *International Journal of Geographical Information Science 22*, 3 (2008), 219–228. doi:10.1080/13658810701626343.

[30] JONES, C. B., PURVES, R. S., CLOUGH, P. D., AND JOHO, H. Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science 22*, 10 (2008), 1045–1065. doi:10.1080/13658810701850547.

[31] JONES, R., ZHANG, W. V., REY, B., JHALA, P., AND STIPP, E. Geographic intention and modification in web search. *International Journal of Geographical Information Science 22*, 3 (2008), 229–246. doi:10.1080/13658810701626186.

[32] KOOPMAN, B., BRUZA, P., SITBON, L., AND LAWLEY, M. Evaluating medical information retrieval. In *Proc. 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2011), ACM, pp. 1139–1140.

[33] LAERE, O. V., SCHOCKAERT, S., TANASESCU, V., DHOEDT, B., AND JONES, C. B. Georeferencing Wikipedia documents using data from social media sources. *ACM Transactions on Information Systems (TOIS) 32*, 3 (2014), 12. doi:10.1145/2629685.

[34] LARSON, R. R., AND FRONTIERA, P. Spatial ranking methods for geographic information retrieval (GIR) in digital libraries. In *ECDL* (2004), R. Heery and L. Lyon, Eds., vol. 3232 of *Lecture Notes in Computer Science*, Springer, pp. 45–56. doi:10.1007/978-3-540-30230-8_5.

[35] LEE, J. H. Analyses of multiple evidence combination. In *Proc. 20th Annual International ACM SIGIR Conference* (New York, NY, 1997), ACM Press, pp. 267–276. doi:10.1145/258525.258587.

[36] LEIDNER, J. L. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Universal Press, Boca Raton, FL, 2007. doi:10.1145/1328964.1328989.

[37] LEVELING, J., AND VEIEL, D. Experiments on the exclusion of metonymic location names from GIR. In *Evaluation of Multilingual and Multi-modal Information Retrieval*, C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. Oard, M. Rijke, and

M. Stempfhuber, Eds., vol. 4730 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2007, pp. 901–904. doi:10.1007/978-3-540-74999-8_114.

[38] LI, Y., MOFFAT, A., STOKES, N., AND CAVEDON, L. Exploring probabilistic toponym resolution for geographical information retrieval. In *Proc. 3rd ACM Workshop On Geographic Information Retrieval (GIR)* (Seattle, WA, 2006), pp. 17–22.

[39] LIEBERMAN, M. D., SAMET, H., SANKARANARAYANAN, J., AND SPERLING, J. STEWARD: Architecture of a spatio-textual search engine. In *Proc. 15th Annual ACM International Symposium on Advances in Geographic Information Systems* (New York, NY, 2007), ACM, pp. 1–8. doi:10.1145/1341012.1341045.

[40] MANDL, T., CARVALHO, P., NUNZIO, G. M. D., GEY, F. C., LARSON, R. R., SANTOS, D., AND WOMSER-HACKER, C. GeoCLEF 2008: The CLEF 2008 cross-language geographic information retrieval track overview. In *Evaluating Systems for Multilingual and Multimodal Information Access* (2008), vol. 5706 of *Lecture Notes in Computer Science*, Springer, pp. 808–821. doi:10.1007/978-3-642-04447-2_106.

[41] MANDL, T., GEY, F. C., NUNZIO, G. M. D., FERRO, N., LARSON, R., SANDERSON, M., SANTOS, D., WOMSER-HACKER, C., AND XIE, X. GeoCLEF 2007: The CLEF 2007 cross-language geographic information retrieval track overview. In *CLEF* (Berlin, 2007), C. Peters, V. Jijkoun, T. Mandl, H. Muller, D. W. Oard, A. Penas, V. Petras, and D. Santos, Eds., vol. 5152 of *Lecture Notes in Computer Science*, Springer, pp. 745–772.

[42] MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. *Introduction to Information Retrieval*. Cambridge University Press, New York, 2008. doi:10.1017/CBO9780511809071.

[43] MARTINS, B., BORBINHA, J., PEDROSA, G., GIL, J. A., AND FREIRE, N. Geographically-aware information retrieval for collections of digitized historical maps. In *Proc. 4th ACM Workshop on Geographical Information Retrieval (GIR)* (2007), ACM, pp. 39–42. doi:10.1145/1316948.1316959.

[44] MARTINS, B., AND CALADO, P. Learning to rank for geographic information retrieval. In *Proc. 6th Workshop on Geographic Information Retrieval (GIR)* (New York, NY, 2010), ACM. doi:10.1145/1722080.1722107.

[45] MOFFAT, A., AND ZOBEL, J. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems 27*, 1 (2008), 2:1–2:27. doi:10.1145/1416950.1416952.

[46] MONTAGUE, M., AND ASLAM, J. A. Condorcet fusion for improved retrieval. In *Proc. 11th International Conference on Information and Knowledge Management (CIKM)* (New York, NY, 2002), ACM, pp. 538–548. doi:10.1145/584792.584881.

[47] NAVIGLI, R. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR) 41*, 2 (2009), 10. doi:10.1145/1459352.1459355.

[48] NEDAS, K. A., AND EGENHOFER, M. J. Spatial-scene similarity queries. *Transactions in GIS 12*, 6 (2008), 661–681. doi:10.1111/j.1467-9671.2008.01127.x.

[49] NIELSEN, J. Participation inequality: Encouraging more users to contribute. http://http://www.nngroup.com/articles/participation-inequality/, 2006.

[50] OSTERMANN, F. O., TOMKO, M., AND PURVES, R. User evaluation of automatically generated keywords and toponyms for geo-referenced images. *Journal of the American Society for Information Science and Technology 64*, 3 (2013), 480–499. doi:10.1002/asi.22738.

[51] PALACIO, D., CABANAC, G., SALLABERRY, C., AND HUBERT, G. On the evaluation of geographic information retrieval systems. *International Journal on Digital Libraries 11*, 2 (2010), 91–109. doi:10.1007/s00799-011-0070-z.

[52] PALACIO, D., DERUNGS, C., AND PURVES, R. Creating test collections from user generated content for GIR evaluation. In *Proc. 7th Workshop on Geographic Information Retrieval (GIR)* (New York, NY, 2013), ACM, pp. 82–83. doi:10.1145/2533888.2533934.

[53] PURVES, R. S., CLOUGH, P., JONES, C. B., ARAMPATZIS, A., BUCHER, B., FINCH, D., FU, G., JOHO, H., SYED, A. K., VAID, S., AND YANG, B. The design and implementation of SPIRIT: A spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science 21*, 7 (2007), 717–745. doi:10.1080/13658810601169840.

[54] RATTENBURY, T., AND NAAMAN, M. Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web 3*, 1 (2009), 1–30. doi:10.1145/1462148.1462149.

[55] ROBERTSON, S., AND ZARAGOZA, H. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009. doi:10.1561/1500000019.

[56] SANDERSON, M. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc, 2010. doi:10.1561/1500000009.

[57] SANDERSON, M., AND KOHLER, J. Analyzing geographic queries. In *Proc. Workshop on Geographic Information Retrieval* (2004).

[58] SANDERSON, M., PARAMITA, M. L., CLOUGH, P., AND KANOULAS, E. Do user preferences and evaluation measures line up? In *Proc. 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, 2010), ACM, pp. 555–562. doi:10.1145/1835449.1835542.

[59] STOKES, N., LI, Y., MOFFAT, A., AND RONG, J. An empirical study of the effects of NLP components on geographic IR performance. *International Journal of Geographical Information Science 22*, 3 (2008), 247–264. doi:10.1080/13658810701626210.

[60] VAID, S., JONES, C. B., JOHO, H., AND SANDERSON, M. Spatio-textual indexing for geographical search on the Web. In *Proc. 9th International Symposium on Advances in Spatial and Temporal Databases (SSTD)* (Berlin, 2005), C. B. Medeiros, M. J. Egenhofer, and E. Bertino, Eds., vol. 3633 of *Lecture Notes in Computer Science*, Springer, pp. 218–235. doi:10.1007/11535331_13.

[61] VOORHEES, E. M. The philosophy of information retrieval evaluation. In *CLEF'01: Proc. Second Workshop of the Cross-Language Evaluation Forum* (2002), C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, Eds., vol. 2406 of *Lecture Notes in Computer Science*, Springer, pp. 355–370. doi:10.1007/3-540-45691-0_34.

[62] VOORHEES, E. M., AND HARMAN, D. K. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA, 2005. doi:10.1162/coli.2006.32.4.563.

[63] WHITE, R., AND BUSCHER, G. Characterizing local interests and local knowledge. In *Proc. SIGCHI Conference on Human Factors in Computing Systems* (Austin, TX, 2012), ACM, pp. 1607–1610. doi:10.1145/2207676.2208283.

[64] WING, B. P., AND BALDRIDGE, J. Simple supervised document geolocation with geodesic grids. In *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Stroudsburg, PA, 2011), vol. 1, Association for Computational Linguistics, pp. 955–964.

[65] YIN, J., KARIMI, S., AND LINGAD, J. Pinpointing locational focus in microblogs. In *Proc. 2014 Australasian Document Computing Symposium (ADCS)* (New York, NY, 2014), ACM, pp. 66:66–66:72. doi:10.1145/2682862.2682868.

[66] ZHANG, W., AND GELERNTER, J. Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science*, 9 (2015), 37–70. doi:10.5311/JOSIS.2014.9.170.